Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models *

DL Oberski[1] A Kirchner[2,3] S Eckman[2] F Kreuter[4,5,6]

[1] Utrecht University, The Netherlands

[2] RTI International, United States

[3] University of Nebraska, United States

[4] Institute for Employment Research, Germany

[5] University of Mannheim, Germany

[6] University of Maryland, United States

Abstract

Administrative data are increasingly important in statistics, but, like other types of data, may contain measurement errors. To prevent such errors from invalidating analyses of scientific interest, it is therefore essential to estimate the extent of measurement errors in administrative data. Currently, however, most approaches to evaluate such errors involve either prohibitively expensive audits or comparison with a survey that is assumed perfect.

We introduce the "generalized multitrait-multimethod" (GMTMM) model, which can be seen as a general framework for evaluating the quality of admin-

*The authors are indebted to Hal Stern and Jörg Drechsler for their comments as well as Barbara Felderer for her assistance in preparing the data. This work was supported by the Netherlands Organization for Scientific Research (NWO) [Veni grant number 451-14-017].

istrative and survey data simultaneously. This framework allows both survey and administrative data to contain random and systematic measurement errors. Moreover, it accommodates common features of administrative data such as discreteness, nonlinearity, and nonnormality, improving similar existing models. The use of the GMTMM model is demonstrated by application to linked survey-administrative data from the German Federal Employment Agency on income from of employment, and a simulation study evaluates the estimates obtained and their robustness to model misspecification.

KEY WORDS: Measurement error, Latent Variable Models, Official statistics, Register data, Administrative data, Reliability

1. INTRODUCTION

Register data and administrative records play an increasingly important role in statistics (Wallgren and Wallgren, 2007) and policy (see, for example, the Commission on Evidence-based Policymaking[1]), and several authors recommend and predict the increased use of "big data" (Entwisle and Elias, 2013; Podesta, 2014), including administrative register data (Japec et al., 2015). Uses to date include studies of how agricultural households affect land changes (Rindfuss et al., 2004), voter turnout (Ansolabehere and Hersh, 2012), or how peoples' numerical ability relates to mortgage default (Gerardi et al., 2013). However, there is evidence that register data may contain considerable measurement errors (Groen, 2012). For example, Bakker (2012, p. 15) estimated that 24% of the variance in Dutch official hourly wages records was random measurement error, and Ladouceur et al. (2007, p. 275) suggested that 20% to 30% of osteoarthritis cases are not registered in Quebec hospital administrative records, causing bias in prevalence estimates. The measurement error present in administrative records can severely bias and invalidate research results (Carroll et al.,

[1] https://cep.gov/

2006; Saris and Gallhofer, 2007; Vermunt, 2010). It is therefore essential to evaluate the extent of measurement error in register data.[2]

The difficulty in studying error in register and administative data, however, is that there is often no “gold standard” measure. Some authors have suggested to link administrative registers to a survey, assuming the survey contains no measurement error (e.g. Yucel and Zaslavsky, 2005). But measurement error in survey data is widespread (Hansen et al., 1961, 1964; Felligi, 1964; Andrews, 1984; Alwin, 2007; Saris and Gallhofer, 2007; Biemer, 2011), and is in fact often measured by taking administrative records as the “gold standard” (e.g. Kapteyn and Ypma, 2007; Kreuter et al., 2010; Sakshaug et al., 2010; Kim and Tamborini, 2014). Thus, we often have two data sources, both measured with error, and we are interested in estimating the error in both.

Very few studies have attempted to estimate measurement error in both survey and administrative data simultaneously. Nordberg et al. (2004) discussed a longitudinal latent Markov model of measurement error in income, but again assumed the administrative register to be perfect in cross-sectional data; Pavlopoulos and Vermunt (2015) applied a similar latent Markov model to unemployment data; and Bakker (2012) and Scholtus et al. (2015) estimated measurement error using linear factor analysis. However, the models used in these studies have several drawbacks when applied to administrative register data. First, true values of the variables of interest are often censored, zero-inflated, gamma, count, or nominal, and thus models which assume normally distributed true values are not appropriate. For example, income is usually zero-inflated and occupation is nominal. Second, the measurement error process in registers is likely to lead to nonnormal and nonlinear errors, yet many models used to study measurement error assume linear and homoskedastic errors.

[2] We use the terms “register data” and “administrative data” synonomously to avoid repetition.

For example, top-coding of income causes nonlinear method effects (Gottschalk and Huynh, 2010), and it is often thought that low earners over-report while high earners under-report, yielding "mean-reverting" random errors (e.g. Kim and Tamborini, 2014). Third, the measurement quality of administrative data often differs over observations, yielding a mixture of measurement models. For example, the records may be obtained from a mixture of sources (Wallgren and Wallgren, 2007), such as both employer statements and employee self-reports, or the variable may be more ambiguously defined for some cases than for others: the income of day laborers is an example. Earlier approaches have not accounted for such heterogeneity. Currently, then, there is no generally applicable method to evaluate the extent of measurement error in register and survey data.

Our contributions to the literature are threefold: first, we present a framework for simultaneously estimating measurement error in register and survey data that addresses the shortcomings of earlier methods. Second, we evaluate the finite sample performance of this model, as well as its robustness to misspecification of key assumption. Third, we apply this framework to a important official register from the German Federal Employment agency. Section 2 introduces the modeling framework used to estimate the extent of measurement error in survey and register data simultaneously. Section 3 evaluates robustness of the model to misspecification, while 4 evaluates its finite sample performance. Section 5 applies the model to linked survey-register data on income of employment from the German Federal Employment agency.

2. MEASUREMENT ERROR ESTIMATION FROM MULTIPLE ERROR-PRONE SOURCES

Measurement error in surveys has been extensively studied, and is often thought to stem from response, coding, processing, and interviewer errors in the data collection process (see Groves and Lyberg, 2010; Biemer et al., 2017). Differences across respondents in the size of these errors will emerge as random noise in the observed variables. Moreover, because different survey variables are usually reported by the same respondent, distinct variables tend to share common errors, a phenomenon known as "method effect" in the literature (Andrews, 1984).

Because surveys contain both random and correlated errors, Andrews (1984) adapted the "multitrait-multimethod" (MTMM) design (Campbell and Fiske, 1959) to survey measurement error estimation. The MTMM design can be described as a within-person experimental design crossing "traits" of interest with measurement "methods". To estimate survey measurement error, Andrews and subsequent authors identified "traits" with survey questions, and "methods" with variations of these questions such as response scales (Saris and Gallhofer, 2007).

This approach has led to a large literature on MTMM modeling using confirmatory factor analysis (CFA) or structural equation modeling (SEM) to estimate the degree of random and systematic measurement error in survey data (e.g. Alwin, 1973; Andrews, 1984; Saris and Andrews, 1991; Saris and Gallhofer, 2007). Extensions for ordinal categorical data using the "ordinal factor analysis" model (Muthén, 1983) have also been applied (Oberski et al., 2008). Recently, Oberski et al. (2015) introduced a latent class factor (Vermunt and Magidson, 2004) MTMM model.

Register data errors have been studied extensively in the statistical data editing literature (De Waal et al., 2011). In this field, the primary goal has been to impute values suspected to be erroneous based on contextual information (covariates). Approaches based on Fellegi and Holt (1976) impose tables of edits while making as few changes as possible and leaving the joint distribution intact (Winkler, 1999). Multiple imputation and Bayesian approaches also aim to impute corrected values, but do so based on a model specifying priors on unlikely combinations, and can quantify uncertainty due to edits (Little and Smith, 1987; Ghosh-Dastidar and Schafer, 2003; Kim et al., 2014). Recently, Boeschoten et al. (2016) demonstrated how edit restrictions can be incorporated into a latent class model, merging the latent variable and model-based editing approaches.

In contrast with the goal of correcting records, the goal of estimating the extent of errors in registers has gained interest only recently. Registers are usually created through data entry and therefore also contain response, coding, and processing errors. However, in addition to these errors, administrative registers have been observed to contain errors that occur during the normal course of administration (Groen, 2012). Among these register-specific errors are time lag, definition error, legally motivated ceiling effects, identification error, and harmonization error (Zhang, 2012). Where registers are obtained from the same source, method effects may also occur (Bakker, 2012). Moreover, the resulting relationship between true value and observed register value is often nonlinear, nonnormal, and differing over different administrative units.

The current methods for the estimation of the extent of measurement error in surveys and registers have important drawbacks, which we address in this study. For survey error estimation, the identification of "methods" with question design features implies that other sources of error aside from these specific design features are uncorrelated. For register error estimation, existing MTMM models lack the nonlinearity, nonnormality, and error process heterogeneity needed to realistically model measurement error in administrative registers. Furthermore, the statistical data editing methods, while useful for imputation, do not estimate the extent of measurement error present in a register. In the following section we address these issues by presenting a novel generalization of the MTMM model.

2.1 The generalized multitrait-multimethod model

Our technique for simultaneously estimating measurement error in survey and administrative data builds on the multitrait-multimethod approach. Instead of identifying "methods" with survey question design, however, we consider the survey and register as "methods". Given a set of variables of interest ("traits") for which observed measurements exist in both the administrative data and a sample survey, our goal is to estimate the degree of measurement error in variables observed in both sources.

Let y_{tm} denote an observed random variable measuring the t-th trait using the m-th method. In the application described here, m will denote either the administrative or the survey measurement.

We use generalized latent variable models (Skrondal and Rabe-Hesketh, 2004) to formulate a measurement model for MTMM data from an administrative register and a survey that can account for non-classical error processes, nonnormal distributions, and categorical data. Generalized latent variable models are built up from (i.) linear GLM predictors; (ii.) GLM links and exponential family distributions; and (iii.) conditional independence relations. To account for possible heterogeneous error process, we additionally include (iv.) a set of mixture components that allow the linear predictors to vary over components.

The conditional independence relations we use result from the MTMM design and are common to all MTMM models, whereas the choice of links and distributions is flexible: for this reason we call our approach a "generalized multitrait-multimethod" (GMTMM) model. The flexibility in links allows us to model nonlinearities and heteroskedasticities in the error process, while the choice of distributions for the latent variables allows for nonnormality of the true values. Finally, the optional finite mixture components allow error processes to differ over units. For example, measures obtained from different administrative databases are likely to have different errors (see Litson et al., 2016, for an example of a continuous-data MTMM model with such mixture components). A mixture component in which no relationship exists between true score and observed score may also be useful in the presence of linkage error (Larsen and Rubin, 2001; Lahiri and Larsen, 2005; Kapteyn and Ypma, 2007).

The main ideas behind the GMTMM model are:

- Observed survey and administrative register values are assumed to originate from a common underlying true value ("trait");
- The relationship between true and observed value is modeled as a GLM regression, allowing for considerable flexibility in nonlinearity and the distribution of measurement errors;
- Conditional on the true values, survey and register values are independent measurements;
- Survey values are mutually dependent, as are register values, allowing for correlated measurement error caused by "method" factors;
- Differential error processes across different types of units can be modeled as a function of unknown mixture components.

We now describe the GMTMM in terms of (i.) the linear GLM predictors, (ii.) the links and distributions, (iii.) the model's conditional independencies, and (iv.) the mixture option for heterogeneous models.

(i.) Linear predictors. For continuous observed data, linear predictors for the observed variables y_{tm} are:

$$\nu_{tm} = \tau_{tm} + \lambda_{tm}\eta_t + \gamma_{tm}\xi_m, \tag{1}$$

where, for identification purposes, the first loading of each trait factor η_t and method factor ξ_m is set to unity, $\lambda_{t1} = \gamma_{1m} = 1$. For categorical observed data, linear predictors for category $y_{tm} = k$ are

$$\nu_{ktm} = \tau_{ktm} + \lambda_{ktm}\eta_t + \gamma_{km}\xi_m, \tag{2}$$

where the first category can be chosen as a reference by setting $\tau_{1tm} = \lambda_{1tm} = \gamma_{1m} = 0$ (e.g. Vermunt and Magidson, 2013).

At times, "paradata" may be available that were captured during the process of survey and register data collection (Kreuter, 2013). Examples for surveys include response times, behavior codes, and vocal pitch; for registers, paradata have not been widely studied, but might include the age of the record or the quality control budget of the department that produced it, if this differs across records. Where such data are informative about the measurement error process, they can be included as covariates in the linear predictor and allowed to interact with the latent "trait" variable. Denoting the paradata covariate as z, the linear predictor then becomes

$$\nu_{tm} = \tau_{tm} + \lambda_{tm}\eta_t + \gamma_{tm}\xi_m + \delta_{tm}z + \beta_{tm}z\eta_t, \tag{3}$$

allowing for both a shift (δ_{tm}) and a different measurement relationship (β_{tm}) across values of the paradata covariate z. To simplify the discussion below, we will omit such covariates from the likelihood.

(ii.) Links and distributions. Each of the observed and latent variables is assigned a distributional "family" and a link function $g(\cdot)$ connecting the linear predictor to the expectation of the response y_{tm} is chosen,

$$g(\mathrm{E}[y_{tm}|\eta_t, \xi_m]) = \nu_{tm}, \qquad \text{or} \qquad g(\mathrm{E}[y_{ktm}|\eta_t, \xi_m]) = \nu_{ktm}, \tag{4}$$

depending on whether the observed variable is continuous or categorical. Different observed variables may be assigned different link functions and distributions.

We denote the choice of the conditional distribution of the observed responses given the latent variables as $f_y := p(y_{tm}|\eta_t, \xi_m)$ with parameter vector $\boldsymbol{\theta}_y$. Similarly, the multivariate distribution of the latent "true score" variables is denoted f_η with parameters $\boldsymbol{\theta}_\eta$ and the distribution of the latent "method" variables f_ξ with parameters $\boldsymbol{\theta}_\xi$. Depending on whether the variables to which they refer are continuous or categorical, f_y, f_ξ and f_η may be probability density or probability mass functions.

A possible extension, which we do not consider here, is to condition the true score distribution f_η on covariates that define edit restrictions (see Boeschoten et al., 2016, for an application of this idea to latent class models). For example, if η represents "married" (1) versus "not married" (0), f_η could be chosen as a logistic regression on a binary covariate "age $<$ 16" (1) versus "age $\geq$ 16" (0), possibly with fixed strongly negative regression coefficient. This would then impose the edit restriction that married persons must be age 16 or above. Estimating the coefficient from data would impose a "soft edit" (De Waal et al., 2012).

(iii.) Conditional independencies. The specification of the homogeneous generalized latent variable model is completed with assumptions of conditional independence that are necessary for identification of the model parameters from observables. These assumptions mirror those of the linear MTMM model.

Assumption 1. The observed variable y_{tm} is conditionally independent of all other observed variables given its trait factor η_t and method factor ξ_m.

Assumption 1 implies that the joint conditional distribution of observed given latent variables can be factored into the univariate conditional distributions, i.e.

$$p(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{t,m} f_y(y_{tm}|\eta_t, \xi_m, \boldsymbol{\theta}_y). \tag{5}$$

Assumption 2. The latent method factors $\boldsymbol{\xi}$ are mutually independent and independent of the trait variables $\boldsymbol{\eta}$.

Assumption 2 implies that the latent variable joint distribution can be factored into

$$p(\boldsymbol{\xi}, \boldsymbol{\eta}|\boldsymbol{\theta}) = f_\eta(\boldsymbol{\eta}|\boldsymbol{\theta}_\eta) \prod_m f_\xi(\xi_m|\boldsymbol{\theta}_\xi). \tag{6}$$

Note that there may still be dependencies among the latent trait variables in the vector $\boldsymbol{\eta}$.

Homogeneous GMTMM likelihood. When the error process is thought to be homogeneous, the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ is

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int\int \left[f_\eta(\boldsymbol{\eta}|\boldsymbol{\theta}_\eta) \prod_m f_\xi(\xi_m|\boldsymbol{\theta}_\xi) \prod_{t,m} f_y(y_{tm}|\eta_t, \xi_m, \boldsymbol{\theta}_y) \right] d\boldsymbol{\eta} d\boldsymbol{\xi}. \tag{7}$$

where assumptions 1 and 2 are used and the integral is defined as a sum for discrete latent variable distributions.

(iv.) Heterogeneous error processes. For heterogeneous error processes, in which a mixture of error processes is thought to be present, define $p(\boldsymbol{y}|S, \boldsymbol{\theta}_s)$ as the component-specific marginal likelihood, with component specific parameters $\boldsymbol{\theta}_s$.

Typically, it is the measurement parameters that are thought to differ over components: that is, the linear predictors are given an additional subscript $\nu_{tm,s}$.

An example of heterogeneous error results from linkage error: similar to the regression model suggested by Lahiri and Larsen (2005), in mislinked records the register would be unrelated to the true value of the survey respondent, which can be modeled by specifying a two-component mixture with $\lambda_{tm,2} = 0$. Another example occurs when administrative delays occur for some units but not others, so that $\tau_{tm,s}$, $\gamma_{tm,s}$, and $\lambda_{tm,s}$ differ. The number of mixture components may be selected using standard methods such as comparison of BIC or AIC (McLachlan and Peel, 2004).

To model such differences, we introduce an unobserved discrete variable S with categories equal to the number of components, so that the marginal likelihood of the observed data becomes

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{S} p(S)p(\boldsymbol{y}|S, \boldsymbol{\theta}_s). \tag{8}$$

Since the mixture proportions $p(S)$ are typically unknown, this implies an additional $|S|$ parameters in $\boldsymbol{\theta}$ to be estimated.

2.2 Identification and estimation of GMTMM models

Consistent estimates of the parameters $\boldsymbol{\theta}$ can be obtained from observations on three traits from linked survey-register data when these parameter are identifiable. The appendix shows that all parameters of the homogeneous GMTMM model are locally identifiable almost everywhere in the parameter space (see Allman et al., 2009) under mild assumptions, given linked survey and register measures of three variables ("traits").

For heterogeneous error processes, identifiability remains an open problem analytically. Even when local identifiability does occur almost everywhere, in practical applications the information matrix can be observed to approach singularity ("empirical underidentification"; Kenny and Kashy, 1992). Specifically, this occurs when the maximum of the likelihood lies close to a point that violates one of the assumptions outlined in the appendix. Considering this issue and pending analytical results for heterogeneous GMTMM models, we suggest to verify (1) empirical identification on data at the converged solution by examining the rank of the information matrix numerically, and (2) following Forcina (2008), to verify invertibility of the information matrix numerically at a large number of random parameter values.

Standard estimation procedures for generalized latent variable models can be used to estimate the GMTMM model (e.g. Skrondal and Rabe-Hesketh, 2004, chapter 6). The most general is to use standard optimization algorithms to maximize the marginal likelihood from Equation 7 or 8. For certain models, such as latent class MTMM models, direct maximization of the marginal likelihood may become unstable. An expectation-maximization (EM) algorithm (McLachlan and Krishnan, 2007) or MCMC sampling of latent variables and parameters can be used by considering the latent variables $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and S to be missing data.

Certain special cases of GMTMM models, including the examples given above, can be estimated using standard software for latent variable modeling such as Latent Gold (Vermunt and Magidson, 2013) or GLAMM (Rabe-Hesketh et al., 2004), that implement this estimation strategy. Moreover, specialized efficient estimation procedures already exist for certain special cases of the GMTMM model. For example, the linear factor analysis MTMM model can be formulated as a covariance structure model with a closed-form marginal likelihood (Bollen, 1989). The ordinal factor analysis (cumulative probit) model can be similarly dealt with by first computing polychoric correlation coefficients (Muthén, 1983). Such models can be fit using standard software for structural equation modeling. Other possible combinations of choices may require specialized software, or can be implemented in general-purpose software such as Stan (Carpenter et al., 2017). An example of a GMTMM model that requires such additional effort is provided in the online supplement with accompanying Stan code. An example of a GMTMM model that requires such additional effort is provided in the online supplement with accompanying Stan code.

This section introduced a generalized multitrait-multimethod model that can be used to estimate measurement error when at least two separate measures of at least three different phenomena are available. The GMTMM model can deal with nonnormality of true values, nonlinearity and heteroskedasticity of errors, and the existence of unknown groups that exhibit differential measurement error. It is therefore applicable to estimating measurement error in administrative register data and surveys simultaneously. It is also more generally applicable to situations where such error structures are thought to exist in multiple error-prone sources.

3. ASYMPTOTIC ROBUSTNESS TO MISSPECIFICATION

The GMTMM model relies on assumptions of independence. When these assumptions are violated, an important question is the extent to which such violations affect the estimates. This section therefore studies the asymptotic sensitivity of GMTMM model estimates to misspecifications of the independence assumptions.

To study robustness, we examine the asymptotic bias of a misspecified GMTMM model. Two key assumptions are examined: (1) the assumption that traits and methods are marginally independent; and (2) the assumption that methods are mutually marginally independent. We study true models, $\mathcal{M}$, say, that violate these assumptions to different degrees, and examine how much asymptotic bias occurs under misspecified models, $\underline{\mathcal{M}}$, that make both assumptions. Following Kuha and Moustaki (2015), this asymptotic bias can be obtained by maximizing the expectation of the misspecified likelihood, $p_{\underline{\mathcal{M}}}$, under the correct model $\mathcal{M}$,

$$\hat{\boldsymbol{\theta}}_{\underline{\mathcal{M}}} = \arg\max_{\boldsymbol{\theta}} E_{\mathcal{M}} \left[p_{\underline{\mathcal{M}}}(\boldsymbol{y}|\boldsymbol{\theta}) \right]. \tag{9}$$

We accomplish this by studying a fully categorical three-trait, two-method GMTMM model, in which all traits, methods, and observed variables are binary variables. This specification is convenient for two reasons. First, as argued by Allman et al. (2009), properties of discrete latent variable models will generalize approximately to continuous-data models. Second, the binary formulation of the model makes it feasible to maximize Equation 9 by enumerating all possible response patterns of $\boldsymbol{y}$ and their expectation under the true model. After calculating these for each condition, a misspecified GMTMM model is fit using expectation-maximization to the true-model expected proportions (see also Rotnitzky and Wypij, 1994; Heagerty and Kurland, 2001; Biemer, 2011, who use a similar approach to study sensitivity to misspecification in other types of models).

The GMTMM model we study has six binary indicators with a logistic link function,

$$P(Y_{tm} = 1|\eta_t, \xi_m) = \left[1 + \exp(\tau_{tm} + \lambda_{tm}\eta_t + \gamma_m \xi_m)\right]^{-1}. \tag{10}$$

The latent variables themselves are binary variables with a multinomial distribution parameterized using the log-linear model

$$P(\eta_1 = k_1, \eta_2 = k_2, \eta_3 = k_3, \xi_1 = l_1, \xi_2 = l_2) = \frac{\exp\left(\mu_{k_1 k_2 k_3 l_1 l_2}\right)}{\sum_{k'_1 k'_2 k'_3 l'_1 l'_2} \exp\left(\mu_{k'_1 k'_2 k'_3 l'_1 l'_2}\right)}, \tag{11}$$

where

$$
\begin{aligned}
\mu_{k_1k_2k_3l_1l_2} = \sum_{t=1}^{3}\alpha_{tk_t} + \sum_{m=1}^{2}\kappa_{ml_t} + \\
\phi_{12}\eta_{1,k_1}\eta_{2,k_2} + \phi_{13}\eta_{1,k_1}\eta_{3,k_3} + \phi_{23}\eta_{2,k_2}\eta_{3,k_3} + \\
\psi^{(tm)}_{11}\eta_{1,k_1}\xi_{1,l_1} + \psi^{(tm)}_{21}\eta_{2,k_2}\xi_{1,l_1} + \psi^{(tm)}_{31}\eta_{3,k_3}\xi_{1,l_1} + \\
\psi^{(tm)}_{12}\eta_{1,k_1}\xi_{2,l_2} + \psi^{(tm)}_{22}\eta_{2,k_2}\xi_{2,l_2} + \psi^{(tm)}_{32}\eta_{3,k_3}\xi_{2,l_2} + \psi^{(mm)}_{12}\xi_{1,k_1}\xi_{2,k_2},
\end{aligned}
$$

with the latent variables "effects-coded" as $\eta_t, \xi_m \in \{-1, +1\}$. In this loglinear model, the first two lines, involving the latent variable loglinear intercepts α and trait-trait dependencies $\phi_{tt'}$ correspond to parameters of the standard GMTMM model. The last two lines contain additional parameters $\psi^{(tm)}$, corresponding to trait-method dependencies, and $\psi^{(mm)}$, corresponding to method-method dependencies. In a standard GMTMM model, these parameters would be set to zero, corresponding to the assumptions of trait-method independence and method-method independence.

To study sensitivity of the parameter estimates to misspecification of trait-method $\psi^{(tm)}$ and method-method dependency $\psi^{(mm)}$, we vary the following factors:

- The size of the loglinear trait-method dependencies:

  $\psi^{(tm)} \in \{-1, -0.5, -0.2, 0, 0.2, 0.5, 1\}$;

- The size of the loglinear method-method dependency:

  $\psi^{(mm)} \in \{-1, -0.5, -0.2, 0, 0.2, 0.5, 1\}$;

- The size of the loglinear trait slope: $\lambda_{tm} \in \{1, 2, 4\}$;

- The size of the loglinear method slope: $\gamma_{tm} \in \{0.5, 1.0\}$.

All loglinear intercepts are set to zero. The trait-trait loglinear dependencies are set to $\phi_{12} = -2$, $\phi_{23} = 2$, $\phi_{13} = 1$.

		Mean Square for Bias in...		
	df	λ_{tm}	γ_{tm}	$\phi_{tt'}$
$\psi^{(tm)}$	6	7.63	360	29.4
$\psi^{(mm)}$	6	0.230	11.3	7.05
λ_{tm}	2	0.848	636	2.86
γ_{tm}	1	0.121	549	2.35
$\psi^{(tm)} \times \lambda_{tm}$	12	2.19	255	5.76
$\psi^{(tm)} \times \gamma_{tm}$	6	0.475	195	5.12
$\psi^{(mm)} \times \lambda_{tm}$	12	0.119	6.08	5.56
$\psi^{(mm)} \times \gamma_{tm}$	6	0.215	6.22	5.40
Residuals	1712	0.116	7.66	1.08

Table 1: ANOVA with main effects and second-order interactions. The Trait-Method dependency $\psi^{(tm)}$ and trait loading λ_{tm} account for the largest mean square.

The parameter values of the true models $\mathcal{M}$ were chosen to correspond to a very wide range of plausible situations. For example, the setting $\lambda_{tm} = 1$ corresponds to an approximate reliability (Pearson correlation between observed variable and trait) of 0.50, while the highest setting $\lambda_{tm} = 4$ corresponds to a reliability of about 0.96. The reliability therefore varies from terrible to excellent. Similarly, the method effect expressed as a Pearson correlation varies between zero and 0.2, which was indicated to be a commonly encountered situation in continuous data by Saris and Gallhofer (2007). For trait-method and method-method dependencies less guidance is available, but it appears plausible that such dependencies, when present, would not be much stronger than the dependencies among the substantive latent variables. The chosen range $(-1, 1)$ can maximally shift a probability by about 0.5, which appears to be a reasonably strong dependency.

Crossing all factors yields a $7 \times 7 \times 3 \times 2$ full factorial design with 294 conditions. For each condition, we generate the expected proportions under the true model $p_{\mathcal{M}}(\boldsymbol{y}|\boldsymbol{\theta})$ and maximize the likelihood of the misspecified model $\underline{\mathcal{M}}$, which incorrectly assumes trait-method and method-method independence, yielding biased parameter values. The asymptotic bias is then defined as the difference between these values and the true values. The outcomes of interest are asymptotic bias in (1) the trait slopes λ_{tm}, (2) the method slopes γ_{tm}, and (3) the trait-trait dependencies $\phi_{tt'}$.

The full tables of results from all 294 conditions are available in the online appendix. An ANOVA of the bias in each of the three outcomes of interest is shown in Table 1. This table shows mean squares for the bias in the three outcomes of interest, using a model with main effects as well as second-order interactions between the misspecification and loading size factors. This summary demonstrates that the largest deviations in the asymptotic bias are accounted for by the Trait-Method dependency $\psi^{(tm)}$ and trait loading λ_{tm}. GMTMM estimates appear to be most sensitive to these factors and their interaction.

To illustrate the size of the asymptotic bias and demonstrate how misspecification relates to it, Figure 1 plots the true trait-method dependency $\psi^{(tm)}$ against the asymptotic bias for the three main outcomes. The columns of this figure correspond to the three outcomes of interest: respectively, asymptotic bias in the trait-trait dependency, the method slopes, and the trait slopes. The rows correspond to conditions with different values of the trait slope. Each plot in the figure shows a misspecification on the horizontal axis, and the incurred asymptotic bias on the vertical axis. Boxplots show the distribution of the bias, while the solid lines connect the median biases encountered.

Figure 1 demonstrates the effect of incorrectly assuming traits and methods to be independent. As expected, at the points corresponding to correctly specified models (intersections of dotted lines), no bias occurs. However, as misspecification increases in either direction, the parameter estimates of a misspecified model will incur some bias. The figure shows that the trait-trait dependency estimates (first column) are most strongly affected by this misspecification. This bias is attenuated as the measurement becomes better (rows), but still considerable at the most extreme

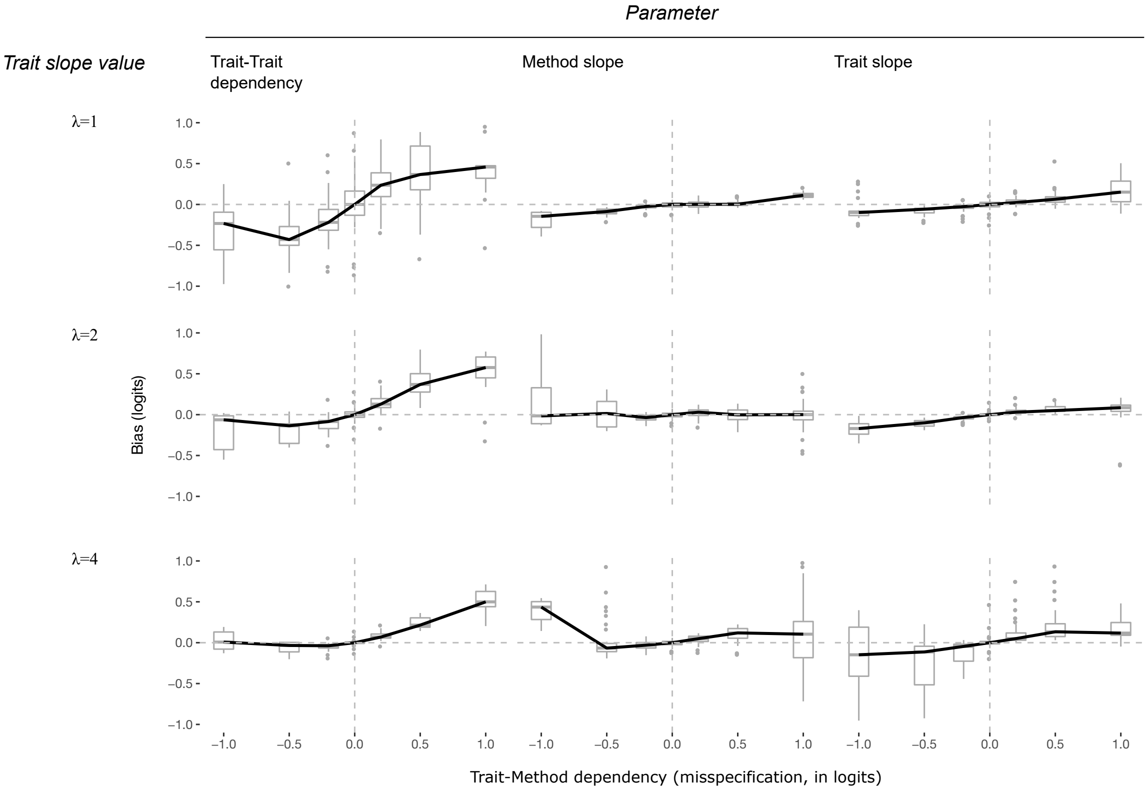

Figure 1: Robustness of the GMTMM model to misspecification of trait-method dependency $\psi^{(tm)}$. Columns show the effect of misspecification on each of three types of parameters. Rows correspond to conditions with different strengths of the trait slope λ_{tm}.

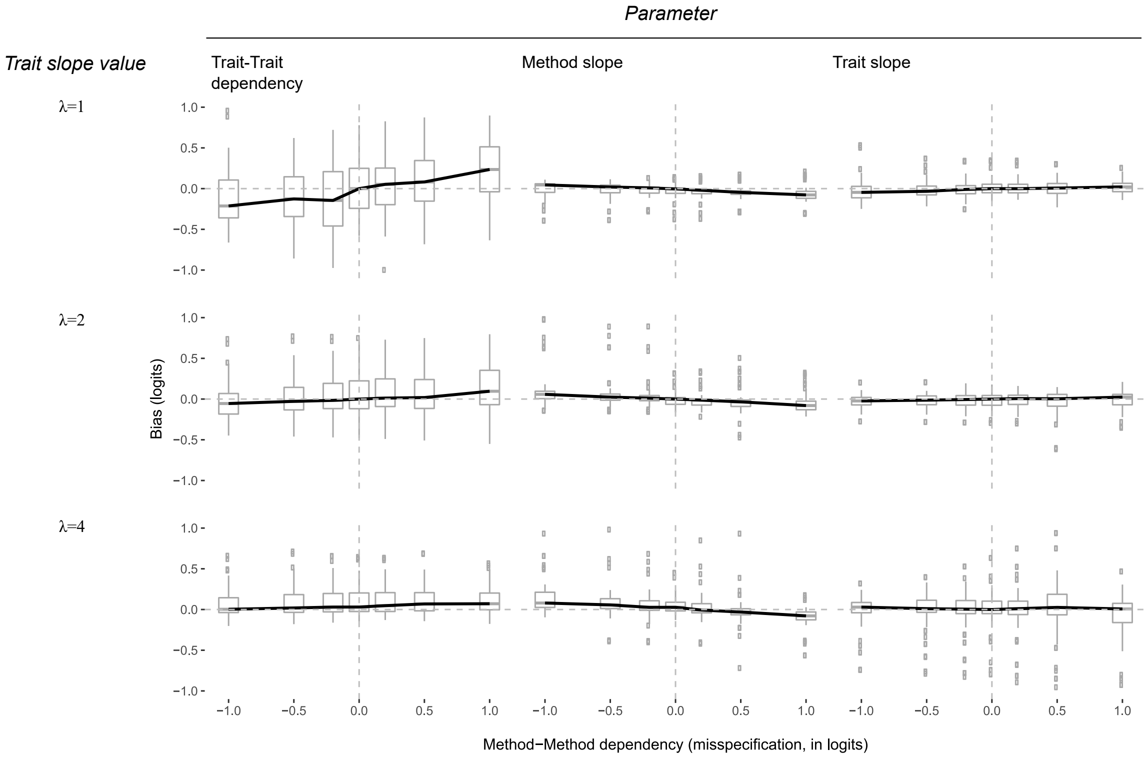

Figure 2: Robustness of the GMTMM model to misspecification of method-method dependency $\psi^{(mm)}$. Columns show the effect of misspecification on each of three types of parameters. Rows correspond to conditions with different strengths of the trait slope λ_{tm}.

values of T-M dependency. The method slopes (second column) are less strongly affected, and appear strongly biased only at the extremes. Finally, the trait slopes, while also affected by this misspecification, do not appear highly sensitive to it.

Figure 2 shows the effect of incorrectly assuming methods to be mutually independent: it plots the same results as Figure 1 as a function of the true method-method dependence, $\psi^{(mm)}$. The flatness of the median bias lines of this figure relative to those in Figure 1 shows that the estimates are rather robust to misspecification of the dependency structure among methods. As the measurement improves (lower rows), this robustness also increases.

The robustness study performed here demonstrates that GMTMM models may be most sensitive to the the assumption of zero trait-method dependency. However, serious biases were only observed at relatively large trait-method dependencies. GMTMM parameter estimates appear to be relatively robust to the assumption of zero method-method dependency. Finally, the parameters of primary interest, the method and trait loadings, were less affected by either type of misspecification than the parameters specifying the joint distribution of true values ("traits").

4. SIMULATION

We demonstrate some key finite sample properties of the maximum likelihood estimates of GMTMM model parameter estimates using a simulation study. Since there are many possible GMTMM models that fall within this framework, we choose a model and parameter values based on our application to linked survey-register dataset obtained from the German Federal Employment Agency, and summarize bias and standard error accuracy under different conditions corresponding to sample sizes.

The response model chosen for the observed variables is a censored regression in which the unobserved trait and method variables are the regressors and the dependent variables are six observed indicators corresponding to the crossing of three traits and two methods. Thus, the response model for the observed variable y_{tm} measuring trait t with method m is

$$y_{tm} = \begin{cases} 0, & \text{if } y^*_{tm} \leq 0 \\ y^*_{tm}, & \text{otherwise} \end{cases}, \tag{12}$$

where y^*_{tm} follows the linear factor model,

$$y^*_{tm} = \tau_{tm} + \lambda_{tm}\eta_t + \gamma_{tm}\xi_m + \epsilon_{tm}, \qquad \epsilon_{tm} \sim N(0, \sigma_{\epsilon,tm}). \tag{13}$$

The latent variables themselves are discrete interval-level variables with a multinomial distribution parameterized using the log-linear model

$$P(\eta_1 = k_1, \eta_2 = k_2, \eta_3 = k_3) = \frac{\exp\left(\mu_{k_1 k_2 k_3}\right)}{\sum_{k_1' k_2' k_3'} \exp\left(\mu_{k_1' k_2' k_3'}\right)}, \tag{14}$$

$$P(\xi_m = k) = \frac{\exp(\kappa_{mk})}{\sum_{k'} \exp(\kappa_{mk'})} \tag{15}$$

where $\mu_{k_1 k_2 k_3} = \sum_{t=1}^{3} \alpha_{t k_t} + \phi_{12}\eta_{1,k_1}\eta_{2,k_2} + \phi_{13}\eta_{1,k_1}\eta_{3,k_3} + \phi_{23}\eta_{2,k_2}\eta_{3,k_3}$.

This model yields the following set of parameters, corresponding to the observed variable intercepts τ_{tm}, trait loadings λ_{tm}, method loadings γ_{tm}, error variances $\sigma_{\epsilon,tm}$, as well as the latent variable loglinear intercepts α_{tk}, and κ_{tk} and latent loglinear associations $\phi_{tt'}$:

$$\boldsymbol{\theta} = (\{\alpha_{tm}\}, \{\kappa_{mk}\}, \{\tau_{tm}\}, \{\lambda_{tm}\}, \{\gamma_{tm}\}, \{\sigma_{\epsilon,tm}\}, \{\phi_{tt'}\})'$$

Furthermore, corresponding to the selected model from our application, we choose three categories for the latent trait and two for the latent method variables:

$$|\eta_t| = 3, |\xi_m| = 2.$$

To ensure parameter values are realistic, we set them to the maximum-likelihood estimates found in our application, and vary the sample size across conditions, $n \in \{200, 500, 1000, 2000\}$. The results of simulating data from this model and analyzing them using the GMTMM model are summarized in Table 2.

Table 2 summarizes the bias, defined as the difference between the true parameter value and the simulation average of the maximum likelihood estimate, as well as the ratio between the average simulation standard error and standard deviation over replications (“s.e./sd”).

It can be seen in Table 2 that under all conditions, the bias is small for most parameters and the estimated standard errors accurately reflect the simulation standard deviation. Exceptions to this good performance are the latent variable intercepts (e.g. α_{21} and κ_{11}) in the condition with the smallest sample size ($n = 200$). Although the bias in this condition is smaller for the other latent intercept parameters, there is a clear pattern of overestimating the size of the largest class and underestimating that of the other classes. This bias disappears as the sample size grows larger. The other parameters do not appear to show any bias, even at the smallest sample size.

Table 2 also shows the performance of information-based standard errors as an estimate of simulation standard deviation. The standard errors perform well when sample size it at least 500. In the smallest sample size condition, some of the standard errors tend to underestimate the simulation standard deviation, which will lead to undercoverage of confidence intervals.

In summary, while the performance of the maximum-likelihood estimates is generally good, bias in some of the parameter estimates and many of the standard errors occurred when the sample size is small ($n = 200$). Therefore, we recommend to use the GMTMM model with samples of at least 500 linked cases.

Table 2: Simulation results for a generalized MTMM model, under different sample sizes. Shown are the true values of the parameters, the simulation bias, and the ratio between the average simulation standard error and standard deviation over replications (“s.e./sd”).

		Sample size n							
		200		500		1000		2000	
Par.	True	Bias	s.e./sd	Bias	s.e./sd	Bias	s.e./sd	Bias	s.e./sd
α_{11}	0.889	0.013	0.956	-0.001	1.002	-0.002	0.968	-0.002	1.013
α_{12}	0.085	-0.009	1.001	0.004	1.088	0.008	1.067	0.004	0.994
α_{21}	1.426	0.074	0.875	0.027	0.964	0.015	0.962	0.013	0.965
α_{22}	-0.305	-0.013	0.943	-0.002	0.999	-0.010	1.020	-0.006	0.985
α_{31}	-0.121	0.017	0.996	-0.003	1.040	-0.007	0.960	-0.002	0.955
α_{32}	-0.356	-0.007	0.948	0.008	1.015	0.010	1.021	0.006	1.069
κ_{11}	0.058	0.018	0.752	0.005	0.902	0.005	0.920	0.001	0.939
κ_{21}	-0.888	-0.015	0.917	-0.008	0.967	-0.003	0.940	-0.005	1.001
τ_{11}	1.296	0.001	0.940	0.003	0.963	-0.000	1.042	-0.001	1.013
λ_{11}	3.772	-0.017	0.815	-0.004	0.917	-0.000	0.948	0.007	0.943
γ_{11}	-1.025	-0.007	1.047	-0.003	1.022	-0.004	1.105	-0.002	0.983
τ_{21}	0.693	-0.015	0.943	-0.000	1.049	0.004	1.065	0.003	1.096
λ_{21}	1.546	0.013	0.956	-0.001	1.005	-0.005	1.010	0.002	0.998
γ_{11}	0.043	0.031	0.850	0.008	0.953	-0.000	0.973	-0.003	0.954
τ_{31}	0.366	0.001	0.870	0.000	0.988	-0.000	0.943	-0.000	0.991
λ_{31}	-0.283	-0.001	0.931	-0.000	1.090	0.000	1.032	0.000	1.008
γ_{31}	0.001	-0.001	0.830	-0.001	0.961	-0.000	1.050	-0.000	1.061
τ_{12}	4.811	0.004	1.025	0.000	1.015	0.005	1.014	0.004	0.950
λ_{12}	2.029	0.003	0.929	-0.001	0.988	-0.004	0.992	-0.003	0.987
γ_{12}	-3.169	-0.003	1.026	0.002	1.023	-0.001	1.038	-0.002	0.958
τ_{22}	1.017	0.009	0.915	0.002	0.982	-0.001	0.947	0.002	0.968
λ_{22}	1.964	-0.003	0.981	-0.001	1.020	0.001	0.960	0.002	0.970
γ_{22}	-0.224	-0.002	0.902	0.001	1.019	0.003	0.966	-0.000	0.967
τ_{32}	0.384	0.001	0.959	-0.000	0.945	0.000	0.968	0.001	1.094
λ_{32}	-0.114	-0.002	0.971	-0.000	0.943	-0.000	0.961	-0.001	0.998
γ_{32}	-0.006	-0.001	0.963	-0.001	0.995	-0.000	1.006	-0.001	1.099
ϕ_{12}	2.916	0.067	0.882	0.032	1.001	0.020	0.969	0.009	0.986
ϕ_{13}	-0.992	-0.012	0.895	-0.033	0.950	-0.008	0.912	-0.000	0.997
ϕ_{23}	-0.289	0.059	0.872	0.020	0.986	0.005	1.016	0.012	0.998
$\sigma_{\epsilon,11}$	0.175	0.004	0.771	0.001	0.934	-0.001	1.005	-0.001	0.984
$\sigma_{\epsilon,21}$	0.420	-0.017	0.993	-0.007	0.971	-0.004	1.055	-0.003	1.074
$\sigma_{\epsilon,31}$	0.003	-0.000	0.891	-0.000	1.031	-0.000	0.932	-0.000	0.941
$\sigma_{\epsilon,12}$	0.545	-0.005	1.043	-0.005	0.931	-0.002	0.940	-0.002	0.980
$\sigma_{\epsilon,22}$	0.141	-0.002	1.067	0.001	1.043	-0.000	1.064	0.000	0.954
$\sigma_{\epsilon,32}$	0.015	-0.000	1.030	-0.000	0.993	-0.000	1.039	-0.000	1.081

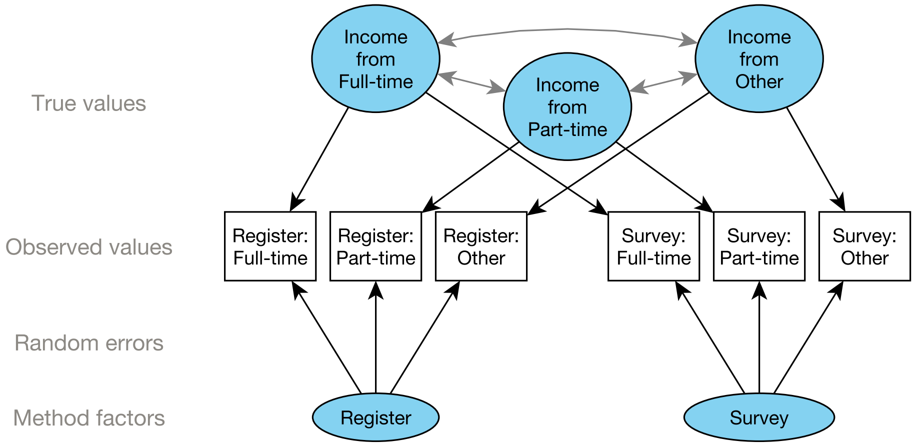

Figure 3: A generalized multitrait-multimethod (GMTMM) model for three "traits" using administrative data and a survey as measurement "methods". The example traits signify personal income from full-time, part-time, and other kinds of employment over a certain period.

5. APPLICATION TO ADMINISTRATIVE DATA ON INCOME

We applied the GMTMM model to a unique dataset provided by the research institute of the German Federal Employment Agency (*Bundesagentur für Arbeit*, BA). The BA's normal operations include job placement and payment of benefits, and for these purposes it maintains an extensive database of citizens' (un)employment histories dating back to 1975. This database covers German employees who are subject to social security contributions as well as recipients of entitlements, comprising about 86% of the overall German labor force. Excluded from the register are most civil servants, the self-employed, and others who have never been in contact with the Agency, such as the never-employed.

Both survey data and the BA's register data are routinely used for labor market and policy research–especially those on income from employment. For consenting respondents, we gained IRB approval to link administrative record data from the Agency with a telephone survey conducted by the Institute for Employment Research (*Institut für Arbeitsmarkt- und Berufsforschung*, IAB). Restricted access to the anonymized linked survey-administrative data was provided at the Agency's offices (IAB Beschäftigtenhistorik (BEH) Version 09.01.00, Nürnberg 2012); the raw data cannot be made publicly available for legal reasons.

Particularly of interest are the BA's records on *income* from full-time, part-time, and "marginal" employment. "Marginal" employment, also known as a "Minijob", is a common form of low-income employment in Germany, yielding monthly income of up to 400 Euro; at or below this maximum, the employee is exempt from income taxes and social security (at the time of data collection).

However, exactly because the income data were collected for the BA's administrative purposes, measurement error can become a serious issue for research in spite of reporting accuracy, because measurement errors in administrative data need not come from the reporting itself (Groen, 2012). For example, although the employers will presumably fulfill their mandate to report accurately, when compiling historical records there may be mismatches and time lapses in an individual's record. Similarly, self-employment periods are absent from the records, again leading to a mismatch in "last part/full-time job", for instance. These issues will lead to random and correlated measurement error for research purposes.

To obtain the survey measurement, a stratified sample of 2,400 respondents was asked to provide information on income from full-time, part-time, and marginal employment (see Eckman et al., 2014, for further description of the sample design). The survey had a response rate (AAPOR RR1) of 19.4%. In the following analyses, we accounted for the sample stratification using complex sampling adjustments. Of the respondents, 2,284 (95%) provided informed consent to record linkage between the survey and the administrative registers. This linkage could be performed using unique person identifiers, so that it seems reasonable to assume no linkage errors were present. By linking the administrative data to the survey data, we thus obtained MTMM designs with three traits and two methods.

The register provides income data only at the level of employment spells. This typically corresponds to an annual basis if a respondent was employed at the same employer throughout a given year. The survey, however, explicitly asks for the last monthly income from gainful employment which is the standard reference period used in most German surveys. Assuming that salaries are paid evenly throughout the employment spell, the administrative data were converted to a monthly basis.

5.1 Estimates of reliability and method effects in survey and administrative measures

To estimate the quality of the administrative register as well as the survey answers on income data, we adapt the model to recognize several aspects of the measurement process:

- Following the econometrics literature (Tobin, 1958), censoring in income is accounted for;

- The relationship between true income and reported income is thought to be nonlinear (Kim and Tamborini, 2014);

- Previous studies linking survey and register data (Scholtus et al., 2015) suggested that there is a subgroup of respondents for whom the two measures correspond exactly, whereas for others they do not, possibly suggesting a heterogeneous error process;

- There is a strong incentive to misreport one's income from a "Minijob" as being equal to or below 400 euros, since at the time of the survey this was the legal maximum income to qualify for tax exemption and social security exemption (see §8 SGB [Social Security Code]).

Due to these factors, a linear Gaussian MTMM will not suffice. Instead, we choose f_y to be the standard censored regression equation, use the "nonparametric" latent class factor analysis formulation of f_ξ and f_η to allow for nonlinearity (Oberski et al., 2015), and investigate whether an additional mixture component of S in which the response is unrelated to the true value fits the data more closely than a homogeneous error structure. This model is no longer a standard structural equation model but can be estimated in the software for latent class (factor) analysis Latent GOLD 5.0 (Vermunt and Magidson, 2013). Program input can be found in the Appendix.

The latent class factor analysis model does not impose a distribution on the latent trait and method factors, but instead approximates these distributions by discrete interval-level latent variables whose category sizes are estimated from the data (Vermunt and Magidson, 2004). Moreover, the possibility of a heterogeneous error structure suggests the presence of an additional discrete nominal latent variable S. Since the number of categories for the latent trait, method, and error structure variables is unknown, we compare the fit of models with differing numbers of categories for each of these. Since increasing the number of categories of the method factors and the error structure variables beyond two never improved the model, we only show these comparisons for models with differing numbers of categories K for the latent trait variables (η_t), with ($|S| = 2$) and without ($|S| = 1$) a heterogenous error structure.

Table 3 shows the fit of these models in terms of loglikelihood (LL), BIC, and AIC, as well as the number of parameters these models have. The model with three latent categories and a heterogeneous error process fit the data best in terms of BIC and AIC. This result suggests that there may indeed be differing error processes for

	Error process							
	Heterogeneous ($\lvert S\rvert = 2$)				Homogeneous ($\lvert S\rvert = 1$)			
K	LL	BIC	AIC	# par.	LL	BIC	AIC	# par.
2	-5060.0	10413.8	10195.9	38	-5388.3	11024.0	10840.6	32
3	-4758.3	**9825.9**	**9596.6**	40	-5272.1	10814.8	10614.1	35
4	-4848.9	10030.3	9783.8	43	-5210.1	10714.1	10496.3	38

Table 3: Fit of GMTMM models for the measurement error in administrative and survey data on income. Rows correspond to models with different numbers of categories K for the latent true score ("trait") variable η_t.

	Trait (λ_{tm})			Method (γ_{tm})		Overall
	1	2	3	1	2	
Administrative data (log-income)						
Full-time	1.11	2.69	4.31			1.85
Part-time	0.65	1.54	2.45			1.08
Marginal	0.09	0.23	0.36			0.21
Survey data (log-income)						
Full-time	2.20	3.16	4.12	5.52	2.25	2.65
Part-time	0.91	1.67	2.45	1.44	1.26	1.28
Marginal	0.27	0.33	0.38	0.33	0.32	0.32

Table 4: Estimated relationships (λ_{tm} and γ_{tm}) between categories of the latent trait variables η and the expected observation of log-income from full-time, part-time, and marginal employment using the administrative and survey measures.

different respondents. Since the model fit did not improve when increasing the number of latent categories from three to four, we selected the three-class heterogeneous model. In other words, we approximate the distribution of true latent income with a discrete three-category latent variable for which the category sizes are estimated. We also allowed for some proportion of the observations to be unrelated to the true value, for example because some fixed value (such as 400 euros) was always chosen in this group regardless of the true income.

Table 4 shows the expected means of the administrative and survey measures of log-income for different categories of the latent trait and method variables. The table illustrates how the observed measures are estimated by the model to relate to the

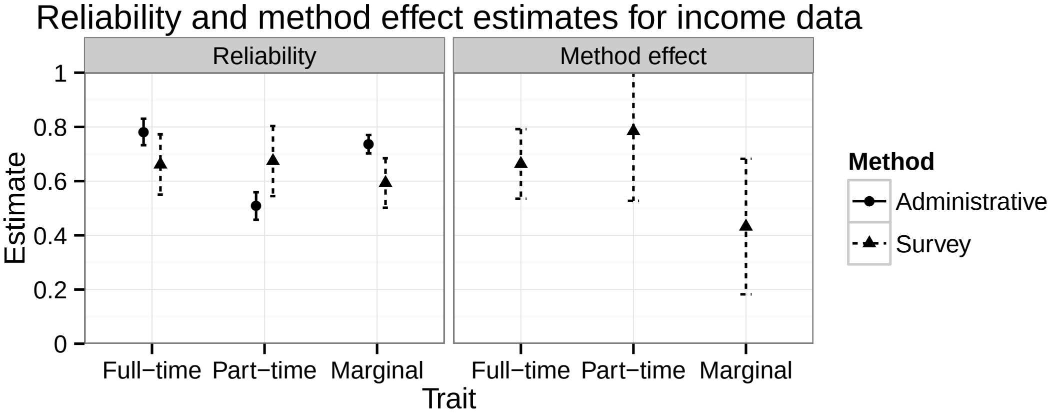

Figure 4: Reliability and method effect estimates for survey data, and reliability estimates for administrative register data on income from full-time, part-time, and "marginal" employment.

respective latent variables. The relationships in Table 4 are marginalized over the two categories of the error process latent variables S. Thus, the table shows how the relationship holds for a respondent whose error process is not known in advance. The estimated proportions of units in each class of S are 0.95 and 0.05. In other words, about 5% (not shown in the table) are estimated to belong to the latent category in which a random value is given – that is, a value that is unrelated to the trait or method variables.

The model is no longer linear, so that reliability and method effect coefficients, which represent (linear) correlations are more difficult to interpret. However, it is possible to calculate the model-implied reliabilities $\text{cor}(y_{tm}, \eta_t)$ and method effects $\text{cor}(y_{tm}, \eta_m)$. These estimates, with confidence intervals based on bootstrapped standard errors, are shown in Figure 4. The figure shows that while the administrative data on income from full-time and marginal jobs are estimated to be superior to the survey measures, the survey measure has a stronger linear correlation with true income level from part-time work. A possible explanation for this difference is a change in mandatory reporting procedures regarding part-time employment in the year 2011. On the other hand, the survey measures do exhibit a strong method dependence, whereas again the administrative register measures were estimated to have no such method dependence.

In summary, we found for official administrative data obtained from the German Federal Employment Agency that the reliability of both survey *and* administrative data was far from perfect. Estimated relationships between these observed variables and other variables of scientific interest will therefore be biased. Moreover, for some of these measures, method effects were found. Such method effects, when ignored, will cause spurious relationships among the true income score ("traits") of interest. When using administrative data, method dependence may be less of a concern. To prevent biases arising from measurement error in substantive analyses of income data, correction methods for known error processes may be needed (e.g. Saris and Gallhofer, 2007; Vermunt, 2010; Skrondal and Kuha, 2012).

6. DISCUSSION AND CONCLUSION

We showed how the quality of survey and administrative data can be evaluated using generalized multitrait-multimethod (GMTMM) models. This approach is an improvement over existing methods, which assume that either the survey or the administrative data are perfect measures. A general framework for data quality evaluation was introduced. This framework is more suited than existing MTMM approaches to administrative data particularities such as categorical measurement, nonlinearities, heterogeneous error processes, and nonnormality. We demonstrated the use of GMTMM models by applying them to administrative and survey data on income of employment from the German Federal Employment Agency. A simulation study demonstrated good properties of the maximum-likelihood estimates for a

GMTMM model with moderate sample sizes, and a robustness study indicated that parameter estimates are not highly sensitive to identifying assumptions.

A clear advantage of our approach is that it allows for the presence of measurement error in both the survey and the administrative register. Furthermore, using the administrative register as a second measure in the MTMM design has an additional advantage over classical MTMM designs using repeated survey measures. When repeated survey measures are used, survey respondents must answer questions on the same topic twice and may remember their answer, creating dependencies that are not modeled (Alwin, 2011), although van Meurs (1995) provided some evidence that this might not occur in practice when sufficient time is allowed between the repetitions. The problem of memory bias does not occur, however, when the measurement methods are administrative and survey data collected separately. Therefore, besides allowing for the estimation of measurement error in administrative records, the MTMM design using linked survey-register data is an attractive method of estimating measurement error in survey variables.

Some limitations of our work remain. First, our model assumed that traits and methods are independent. While the robustness study indicates that the parameters of primary interest may not be highly sensitive to this assumption, it cannot rule out that very strong dependencies between traits and methods will produce bias. We note that it is possible to define a subclass of identifiable GMTMM models that do allow for dependencies among traits and methods, and between methods (linear MTMM models are known to lie outside this subclass, e.g. Kenny and Kashy, 1992). However, this subclass will rely heavily on higher-order moments for identification, which in practice may lead to high-variance estimates. In future studies, it will be of interest to investigate the conditions under which such models can be applied.

Second, we did not discuss model fit evaluation. This issue is not specific to

GMTMM modeling, so that the standard machinery available for global and local fit assessment in generalized latent variable models can be applied to GMTMM modeling (see, e.g. Skrondal and Rabe-Hesketh, 2004; Oberski and Vermunt, 2013; Oberski et al., 2013). Second, little is known about the small sample properties of GMTMM model estimates. While simulation results by Scholtus et al. (2015) on the linear MTMM model were positive, other types of GMTMM models were not evaluated. This remains a topic for future research. Finally, in our application on German data, unique identifiers were available that allowed for close linkage between the survey and register. In other applications, however, such identifiers may not be available for legal reasons or they may not exist. In such cases, linkage error will occur as well as measurement error. As indicated, the heterogenous error process may be employed to model such errors in a fashion similar to the observed multiple regression models of Lahiri and Larsen (2005). However, evaluating the performance of this solution and the interaction between linkage and measurement error remains a topic for future study.

REFERENCES

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Alwin, D. F. (1973). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. *Sociological methodology*, 5:79–105.

Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement.* Wiley-Interscience, New York.

Alwin, D. F. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. In Madans, J., Miller, K., Maitland, A., and Willis, G., editors, *Question Evaluation Methods: Contributing to the Science of Data Quality*, pages 263–293. Wiley Online Library, New York.

Andrews, F. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 48(2):409–442.

Ansolabehere, S. and Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, 20(4):437–459.

Bakker, B. F. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, 66(1):8–17.

Biemer, P. (2011). *Latent Class Analysis of Survey Error.* Wiley, New York.

Biemer, P. P., De Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, C., and West, B. T. (2017). *Total survey error in practice: improving quality in the era of big data.* John Wiley & Sons, New York.

Boeschoten, L., Oberski, D., and de Waal, T. (2016). Estimating classification error under edit restrictions in combined survey-register data. CBS discussion paper 2016/12, Statistics Netherlands, Den Haag. Available from: `https://www.cbs.nl/-/media/_pdf/2016/38/estimating-classification-error-under-edit-restrictions.pdf`.

Bollen, K. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, New York.

Campbell, D. and Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56:81–105.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*, volume 105. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton, FL.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of statistical data editing and imputation*, volume 563. John Wiley & Sons.

De Waal, T., Pannekoek, J., and Scholtus, S. (2012). The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):204–210.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78(3):721–733.

Entwisle, B. and Elias, P. (2013). *New Data for Understanding the Human Condition: International Perspectives.* OECD, Paris, France. Available from: `http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf`.

Fellegi, I. P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical association*, 71(353):17–35.

Felligi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59(308):1016–1041.

Forcina, A. (2008). Identifiability of extended latent class models with individual covariates. *Computational Statistics & Data Analysis*, 52(12):5263–5268.

Gerardi, K., Goette, L., and Meier, S. (2013). Numerical ability predicts mortgage default. *Proceedings of the National Academy of Sciences*, 110(28):11267–11271.

Ghosh-Dastidar, B. and Schafer, J. L. (2003). Multiple edit/multiple imputation for multivariate continuous data. *Journal of the American Statistical Association*, 98(464):807–817.

Gottschalk, P. and Huynh, M. (2010). Are earnings inequality and mobility overstated? the impact of nonclassical measurement error. *The Review of Economics and Statistics*, 92(2):302–315.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics (JOS)*, 28(2).

Groves, R. M. and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879.

Hansen, M., Hurwitz, W., and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38(2):359–374.

Hansen, M., Hurwitz, W., and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. In Rao, C. R., editor, *Contributions to Statistics, Presented to Professor P. C. Mahalanobis on the Occasion of his 70th Birthday*. Pergamon Press, Calcutta, India.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). *AAPOR Report on Big Data*. American Association for Public Opinion Research (AAPOR). Available from: `http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/images/BigDataTaskForceReport_FINAL_2_12_15_b.pdf`.

Kapteyn, A. and Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, 25(3):513–551.

Kenny, D. A. and Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112(1):165.

Kim, C. and Tamborini, C. R. (2014). Response error in earnings an analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research*, 43(1):39–72.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3):375–386.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*. John Wiley & Sons, New York.

Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and measurement error in employment research nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly*, 75(5):pp. 880–906.

Kuha, J. and Moustaki, I. (2015). Nonequivalence of measurement in latent variable modeling of multigroup data: A sensitivity analysis. *Psychological methods*, 20(4):523.

Ladouceur, M., Rahme, E., Pineau, C. A., and Joseph, L. (2007). Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics*, 63(1):272–279.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American statistical association*, 100(469):222–230.

Larsen, M. D. and Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41.

Litson, K., Geiser, C., Burns, G. L., and Servera, M. (2016). Examining trait× method interactions using mixture distribution multitrait–multimethod models. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–21.

Little, R. J. and Smith, P. J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82(397):58–68.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons, New York.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons, New York.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22:43–65.

Nordberg, L., Rendtel, U., and Basic, E. (2004). Measurement error of survey and register income. In Ehling, M. and Rendtel, U., editors, *Harmonisation of Panel Surveys and Data Quality*, pages 65–88. Statistisches Bundesamt, Wiesbaden.

Oberski, D., Saris, W. E., and Hagenaars, J. (2008). Categorization errors and differences in the quality of questions across countries. In Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell, B.-E., and Smith, T. W., editors, *Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC)*. Wiley, New York.

Oberski, D., Van Kollenburg, G., and Vermunt, J. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3).

Oberski, D. and Vermunt, J. (2013). A model-based approach to goodness-of-fit evaluation in item response theory. *Measurement: Interdisciplinary Research & Perspectives*, 11:117–122.

Oberski, D. L., Hagenaars, J. A., and Saris, W. E. (2015). The latent class multitrait-multimethod model. *Psychological methods*, 20(4):422.

Pavlopoulos, D. and Vermunt, J. K. (2015). Measuring temporary employment. do survey or register data tell the truth? *Survey Methodology*, 41:1.

Podesta, J. (2014). Big data: Seizing opportunities preserving values. Technical report, Executive Office of the President. Available from: `https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf`.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2):167–190.

Rindfuss, R. R., Walsh, S. J., Turner, B., Fox, J., and Mishra, V. (2004). Developing a science of land change: challenges and methodological issues. *Proceedings of the*

National Academy of Sciences of the United States of America, 101(39):13976–13981.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, 50(4):1163–1170.

Sakshaug, J. W., Yan, T., and Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection: tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, 74(5):907–933.

Saris, W. E. and Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S., editors, *Measurement errors in surveys*, pages 575–599. John Wiley & Sons, New York.

Saris, W. E. and Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience, New York.

Scholtus, S., Bakker, B. F., and Van Delden, A. (2015). Modelling measurement error to estimate bias in administrative and survey variables. In *New Techniques and Technologies for Statistics (NTTS) conference*, pages 9–13. Available from: `https://www.cbs.nl/-/media/imported/documents/2015/46/modelling_measurement_error.pdf`.

Skrondal, A. and Kuha, J. (2012). Improved regression calibration. *Psychometrika*, 77(4):649–669.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling : multilevel, longitudinal, and structural equation models*. Interdisciplinary statistics series. Chapman & Hall/CRC, Boca Raton, FL.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 26(1):24–36.

van Meurs, L. (1995). Memory effects in MTMM studies. In Saris, W. E. and Münnich, A., editors, *The multitrait-multimethod approach to evaluate measurement instruments*. Eötvös University Press, Budapest.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18:450–469.

Vermunt, J. K. and Magidson, J. (2004). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In van der Ark, L. A., Croon, M. A., and Sijtsma, K., editors, *New developments in categorical data analysis for the social and behavioral sciences*, pages 41–63. Erblaum, Mahwah.

Vermunt, J. K. and Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc., Belmont, MA.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*. Wiley, New York.

Winkler, W. E. (1999). State of statistical data editing and current research problems. Technical report, US Bureau of the Census.

Yucel, R. M. and Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association*, 100(472):1123–1132.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41–63.